BAB III

METODOLOGI PENELITIAN

3.1 Gambaran Umum Objek Penelitian

Penelitian ini bertujuan untuk menganalisis sentimen masyarakat terhadap mobil listrik dengan memanfaatkan komentar pengguna yang terdapat pada platform YouTube. Fokus utama dari studi ini adalah untuk mengidentifikasi persepsi publik terhadap penggunaan mobil listrik sebagai solusi transportasi berkelanjutan. Pengumpulan data dilakukan melalui teknik *scraping* menggunakan bahasa pemrograman Python pada lingkungan Jupyter Notebook, di mana komentar-komentar yang berkaitan dengan mobil listrik diekstraksi dari beberapa video bertema otomotif di YouTube. Analisis sentimen dilakukan untuk mengetahui kecenderungan opini pengguna, apakah bersifat positif, negatif, maupun netral, yang selanjutnya diproses menggunakan algoritma *Support Vector Machine (SVM)*, *Naive Bayes,* dan *Random Forest* sebagai metode klasifikasi.

3.2 Metode Penelitian

Dalam penelitian ini, digunakan metodologi CRISP-DM (Cross-Industry Standard Process for Data Mining) sebagai kerangka kerja untuk menggambarkan siklus hidup proses data mining secara sistematis. Metodologi ini menyediakan panduan umum terhadap tahapan-tahapan proyek data mining, pekerjaan yang dilakukan di setiap tahap, serta hubungan antara aktivitas-aktivitas tersebut. Pendekatan ini dianggap fleksibel dan dapat diterapkan secara luas, baik dalam konteks bisnis maupun akademik, untuk menyelesaikan berbagai permasalahan menggunakan teknik data mining yang sesuai. Penelitian ini mengacu pada enam tahapan utama CRISP-DM, yaitu: *business understanding*, *data understanding*, *data preparation*, *modeling*, *evaluation*, dan *deployment*, yang seluruhnya digunakan untuk mendukung proses analisis dan interpretasi hasil sentimen terhadap mobil listrik.

Algoritma Naive Bayes, Support Vector Machine (SVM), dan Random Forest digunakan sebagai metode klasifikasi dalam penelitian ini, dengan implementasi berbasis bahasa pemrograman Python melalui platform *Jupyter Notebook*. Kedua

algoritma dipilih karena memiliki kinerja yang baik dalam tugas klasifikasi teks, khususnya dalam analisis sentimen.

3.3 Teknik Pengumpulan Data

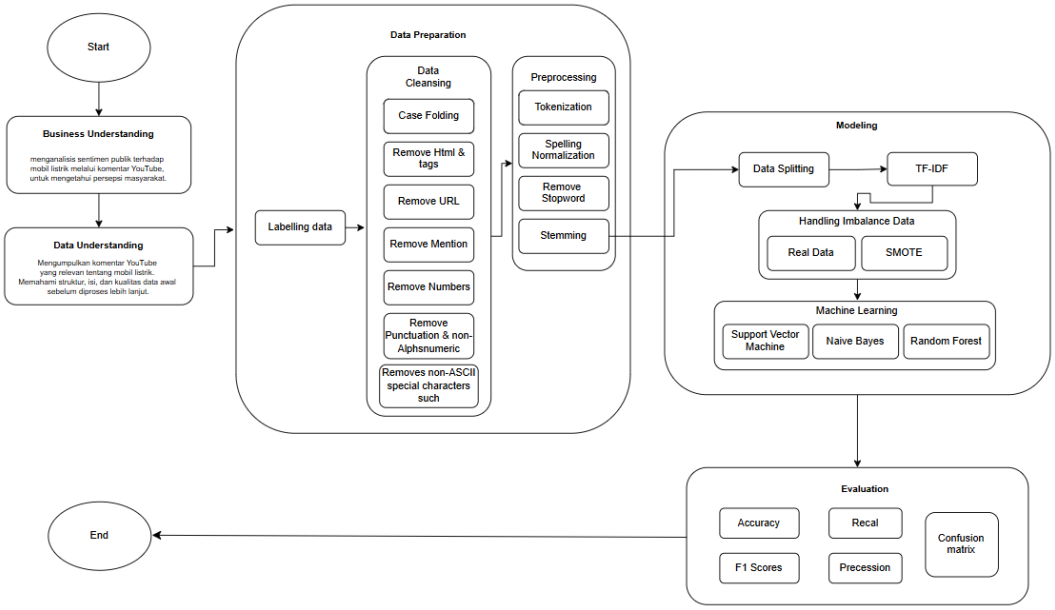
Data yang digunakan dalam penelitian ini berasal dari komentar pengguna pada platform YouTube, yang diperoleh secara langsung melalui proses penarikan data dari server YouTube. Pengumpulan data dilakukan dengan memanfaatkan fitur YouTube API (Application Programming Interface), yang digunakan untuk mengakses dan mengambil komentar dari video tertentu. Dalam prosesnya, penelitian ini menggunakan ID video YouTube sebagai acuan untuk menentukan sumber data. Setelah ID video dimasukkan, sistem akan mengambil seluruh komentar yang tersedia pada video tersebut melalui API dan menyimpannya dalam format file CSV.

Prosedur ini mempermudah pengumpulan data secara otomatis dan terstruktur untuk keperluan analisis sentimen.Seluruh komentar dikumpulkan dalam bentuk file berformat CSV, yang kemudian digunakan sebagai dataset utama dalam pelatihan dan pengujian model. Adapun periode pengambilan data ditentukan selama tahun 2021 hingga 2025, dengan mempertimbangkan video-video yang menampilkan ulasan dan diskusi paling relevan. Pemilihan komentar yang digunakan dalam analisis didasarkan pada indikator keterlibatan pengguna, seperti jumlah likes, panjang komentar, serta kesesuaian konten komentar terhadap topik mobil listrik.

Dalam proses pengumpulan data, peneliti berhasil memperoleh sebanyak 10.000 komentar mentah dari delapan video YouTube yang secara khusus membahas topik mobil listrik dengan kata kunci “mobil listrik,” “EV,” dan “kendaraan listrik.” Pemilihan video dilakukan secara sistematis berdasarkan kriteria tertentu, yaitu channel YouTube yang dikenal melakukan review jujur, jumlah likes pada video, jumlah komentar yang tersedia, serta tahun publikasi video dalam rentang lima tahun terakhir. Langkah ini dilakukan untuk memastikan bahwa data yang dikumpulkan memiliki relevansi yang tinggi dan dapat merepresentasikan opini publik secara lebih akurat terhadap isu mobil listrik yang diteliti.

3.4 Teknik Analisis Data

Pada penelitian ini, digunakan pendekatan model CRISP-DM (Cross-Industry Standard Process for Data Mining) sebagai kerangka kerja dalam proses analisis data. Sebagaimana ditampilkan pada Gambar 3.1, terdapat enam tahapan utama dalam model CRISP-DM yang akan diterapkan secara sistematis dalam penelitian ini. Adapun penjelasan dari masing-masing tahapan disajikan sebagai berikut.



Gambar 3.1 Teknik Analisis Data

3.4.1 Business Understanding

Tahapan *Business Understanding* merupakan langkah awal dalam metodologi CRISP-DM (Cross Industry Standard Process for Data Mining) yang bertujuan untuk merumuskan secara jelas tujuan dan ruang lingkup penelitian. Dalam konteks penelitian ini, tujuan utama adalah untuk menganalisis sentimen masyarakat terhadap mobil listrik berdasarkan komentar yang diperoleh dari platform YouTube. Penelitian ini menggunakan algoritma Naive Bayes dan Support Vector Machine (SVM) untuk mengklasifikasikan opini publik ke dalam kategori sentimen positif, dan negatif. Pemahaman terhadap perspektif masyarakat terhadap mobil listrik menjadi dasar dalam menentukan arah analisis sentimen serta membantu dalam mengidentifikasi pola-pola opini yang berkembang di media sosial.

3.4.2 Data Understanding

Pada tahapan ini, pengambilan data dilakukan dengan menggunakan kata kunci yang berkaitan dengan topik mobil listrik, seperti "mobil listrik", "EV", "kendaraan listrik", dan istilah relevan lainnya. Proses pengumpulan data dilakukan melalui platform YouTube dengan menargetkan komentar-komentar dari video yang membahas isu, ulasan, maupun perbandingan kendaraan listrik. Data dikumpulkan dengan bantuan *Jupyter* melalui proses *scraping* menggunakan bahasa pemrograman Python. Dari proses tersebut, diperoleh sejumlah data komentar yang cukup signifikan untuk digunakan dalam proses analisis sentimen, dengan total data yang memadai untuk mencerminkan opini publik secara umum terhadap mobil listrik.

3.4.3 Data Preparation

Pada tahap ini dilakukan proses pembersihan data yang diawali dengan *cleansing* dan dilanjutkan dengan *text preprocessing*. Proses *text preprocessing* mencakup beberapa tahapan, seperti *case folding*, *tokenizing*, *stemming*, dan *stopword removal*. Seluruh tahapan ini dilakukan untuk memastikan bahwa data yang akan dianalisis memiliki struktur yang bersih dan konsisten, sehingga dapat menghasilkan analisis yang lebih akurat. Tahapan *data preparation* dalam penelitian ini mencakup proses *data cleansing*, *preprocessing*, *labeling*, *TF-IDF*, *SMOTE*, dan *data splitting*.

3.4.3.1 *Labeling Data*

Pada tahap pelabelan data, proses penentuan kategori sentimen terhadap komentar dilakukan secara manual oleh tiga orang relawan. Relawan yang terlibat terdiri dari dua orang pengguna mobil listrik yang telah memiliki pengalaman langsung dalam menggunakan kendaraan tersebut, serta satu orang yang bekerja sebagai mekanik dan memiliki pengetahuan teknis mengenai performa mobil listrik. Ketiga relawan ini secara independen meninjau setiap komentar yang telah dikumpulkan dan memberikan label sentimen yang sesuai berdasarkan konteks isi komentar. Labelling dilakukan oleh ketiga relawan tersebut, di mana setiap komentar dinilai oleh seluruh relawan. Kemudian, label sentimen yang digunakan

untuk setiap komentar ditentukan berdasarkan suara mayoritas dari ketiga penilai. Dalam penelitian ini, kategori sentimen yang digunakan terdiri dari tiga kelas, yaitu positif, negatif, dan netral.

3.4.3.2 *Data Cleansing*

Data cleansing merupakan tahap untuk membersihkan data dari berbagai elemen yang tidak relevan, seperti kesalahan penulisan, duplikasi, maupun informasi yang tidak lengkap. Tujuan dari proses ini adalah untuk memastikan bahwa data yang digunakan dalam analisis memiliki tingkat akurasi dan konsistensi yang tinggi, sehingga hasil analisis menjadi lebih valid dan dapat diandalkan. Pada penelitian ini, proses *data cleansing* dilakukan melalui beberapa tahapan untuk membersihkan data teks dari elemen-elemen yang tidak relevan dan mengganggu analisis. Langkah pertama adalah *case folding*, yaitu mengubah seluruh huruf dalam teks menjadi huruf kecil agar konsisten. Selanjutnya, teks dibersihkan dari elemen HTML dan tag yang tidak diperlukan. Setelah itu, dilakukan penghapusan URL yang sering muncul dalam komentar atau deskripsi, diikuti dengan penghapusan *mention* seperti username yang diawali dengan simbol "@". Tahap berikutnya adalah menghapus angka, karena angka umumnya tidak memiliki kontribusi penting dalam analisis sentimen. Kemudian, tanda baca dan karakter non-alfanumerik dibuang untuk menyederhanakan struktur teks. Terakhir, karakter khusus non-ASCII (American Standard Code for Information Interchange) seperti emoji dan simbol asing juga dihapus untuk memastikan teks dapat diproses dengan baik oleh algoritma pemodelan.

3.4.3.3 *Preprocessing*

Tahapan *preprocessing* merupakan proses awal dalam pengolahan data teks yang bertujuan untuk mempersiapkan data mentah sebelum digunakan dalam proses analisis atau pelatihan model *machine learning*. Proses ini penting dilakukan untuk menyederhanakan dan membersihkan data agar dapat dikenali dan diproses dengan lebih optimal oleh algoritma

klasifikasi. Dalam penelitian ini, preprocessing dilakukan melalui beberapa langkah utama.

4. *tokenization* digunakan untuk memecah kalimat menjadi satuan kata, sehingga setiap kata dapat dianalisis secara terpisah.
5. *Spelling Normalization*, yakni proses standarisasi kata-kata tidak baku atau kata gaul menjadi bentuk baku sesuai KBBI agar seragam dan mudah dianalisis.
6. *stopword removal*, yaitu proses penghapusan kata-kata umum yang tidak memiliki kontribusi penting terhadap makna teks, seperti "dan", "yang", atau "atau".
7. *stemming* untuk mengubah setiap kata ke bentuk dasarnya, misalnya kata "mengelola" menjadi "kelola", sehingga berbagai bentuk turunan kata dapat dianggap sebagai satu entitas yang sama dalam analisis.

3.4.4 Modeling

Pada tahap ini, dilakukan proses pemodelan dengan menggunakan dua algoritma klasifikasi yang telah dipilih, yaitu Naive Bayes dan Support Vector Machines (SVM). Pemilihan kedua algoritma ini didasarkan pada referensi dari berbagai penelitian sebelumnya yang menunjukkan bahwa keduanya memiliki performa yang baik dalam melakukan klasifikasi sentimen terhadap data teks. Implementasi algoritma dilakukan menggunakan bahasa pemrograman Python, yang mendukung berbagai pustaka (*library*) untuk pengolahan teks dan machine learning, sehingga proses pemodelan dapat berjalan secara efisien dan terstruktur. Model yang dibangun akan dilatih menggunakan data yang telah diproses dan diberi label, kemudian dievaluasi untuk mengetahui tingkat akurasi serta kemampuan klasifikasi terhadap komentar mengenai mobil listrik.

3.4.4.1 *Data Splitting*

Pada tahap *data splitting*, data yang telah melalui proses praproses dan pelabelan akan dibagi menjadi dua bagian, yaitu data latih (*training data*) dan data uji (*testing data*). Pemisahan ini bertujuan untuk mengukur

kinerja model dalam memprediksi data yang belum pernah dilihat sebelumnya. Dalam penelitian ini, digunakan rasio pembagian sebesar 80% untuk data *training* dan 20% untuk data *testing*. Rasio ini dipilih karena dinilai mampu memberikan jumlah data pelatihan yang cukup untuk membangun model yang akurat, sekaligus menyediakan data uji yang representatif untuk mengevaluasi performa model secara objektif. Penggunaan rasio 80:20 juga telah banyak digunakan dalam penelitian analisis sentimen dan terbukti efektif dalam mendukung proses klasifikasi berbasis machine learning.

3.4.4.2 *TF-IDF*

Term Frequency-Inverse Document Frequency (TF-IDF) merupakan salah satu metode dalam pengolahan teks yang digunakan untuk mengukur tingkat kepentingan suatu kata dalam sebuah dokumen terhadap keseluruhan koleksi dokumen. Teknik ini memberikan bobot tertentu pada setiap kata berdasarkan seberapa sering kata tersebut muncul dalam satu komentar (term frequency) dan seberapa jarang kata tersebut muncul di seluruh komentar lainnya (inverse document frequency). Tujuan dari penerapan TF-IDF adalah untuk mentransformasikan data teks menjadi bentuk vektor numerik, sehingga dapat digunakan oleh algoritma klasifikasi dalam proses pelatihan dan pengujian model. Dengan demikian, kata-kata yang dianggap lebih penting dalam konteks komentar akan memiliki bobot yang lebih tinggi dibandingkan kata-kata umum yang sering muncul di seluruh dokumen.

3.4.4.3 *Handling Imbalance Data*

Pada tahap penanganan ketidakseimbangan data (Handling Imbalance Data), penelitian ini menerapkan dua pendekatan sebelum proses pelatihan model untuk memitigasi bias akibat dominasi kelas mayoritas. Pertama, Real Data mempertahankan distribusi asli label misalnya 90 % komentar negatif dan 10 % komentar positif sehingga validitas data tetap terjaga, tetapi berisiko membuat model "memfavoritkan" kelas mayoritas). Kedua, SMOTE (Synthetic Minority Over-sampling Technique), membuat

sampel sintetis pada kelas minoritas dengan menghitung vektor antar tetangga terdekat, sehingga proporsi kelas menjadi lebih seimbang tanpa menduplikasi data mentah. Dalam konteks skripsi, pilihan antara kedua skema ini dievaluasi melalui eksperimen komparatif menggunakan metrik precision, recall, dan F1-score untuk menentukan apakah SMOTE dapat meningkatkan deteksi kelas minoritas tanpa menurunkan kemampuan generalisasi model terhadap data nyata.

3.4.4.4 *Machine Learning*

Pada tahap ini, dilakukan proses pemodelan dengan menggunakan dua algoritma klasifikasi yang telah dipilih, yaitu Support Vector Machines (SVM) dan Naive Bayes. Pemilihan kedua algoritma ini didasarkan pada referensi dari berbagai penelitian sebelumnya yang menunjukkan bahwa keduanya memiliki performa yang baik dalam melakukan klasifikasi sentimen terhadap data teks. Implementasi algoritma dilakukan menggunakan bahasa pemrograman Python, yang mendukung berbagai pustaka (library) untuk pengolahan teks dan machine learning, sehingga proses pemodelan dapat berjalan secara efisien dan terstruktur. Model yang dibangun akan dilatih menggunakan data yang telah diproses dan diberi label, kemudian dievaluasi untuk mengetahui tingkat akurasi serta kemampuan klasifikasi terhadap komentar mengenai mobil listrik.

3.4.5 Evaluation

Pada tahap evaluasi, dilakukan pengujian terhadap performa kedua algoritma yang digunakan, yaitu Support Vector Machine (SVM) dan Naïve Bayes, dengan menggunakan beberapa metrik evaluasi seperti accuracy, precision, recall, dan F1-Score. Penggunaan metrik-metrik ini bertujuan untuk mengetahui seberapa baik masing-masing model dalam mengklasifikasikan sentimen komentar pengguna terhadap mobil listrik. Setelah hasil pengujian dianalisis, model dengan performa terbaik akan dipilih sebagai dasar dalam menarik kesimpulan akhir dan memberikan rekomendasi terkait persepsi masyarakat terhadap mobil listrik berdasarkan data yang diperoleh dari komentar YouTube.